Wrangle Report - Project 7 : Wrangle and Analyze Data

The dataset that I wrangled (and analyzed and made vizualizations) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Our goal was to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

My tasks in this project are as follows:

Data wrangling, which consists of:

1.Gathering data

2.Assessing data

3.Cleaning data

Data wrangling

1.Gathering data

From different sources and in 3 different formats, I:

- downloaded manually from the link provided in the classroom : twitter_archive_enhanced.csv
- downloaded programmatically using the Requests library: image_predictions.tsv
- downloaded manually from the attachment provided in the classroom: tweet_json.txt

I stored each one of them in an corresponding dataframe.

2.Assesing data

Assesing data means both visually and programmatically.

Visually assessing the data, refers to looking visually on each dataframe.
Programmatically assessing of data, means using Python.

I identified quality issues and tidiness issues, and I made a summary, as follows.

Qualiy issues:

1. 'id' column name from the tweet_json_df dataframe, needs to be changed to 'tweet_id'
2. unnecessary columns with probability and algorithms used, containing non-descriptive columns : 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog', 'img_num', from image_predictions_df dataframe

3. non-descriptive columns for dog breed 'p1' and 'p1_conf'; rename 'p1', 'p1_conf' to 'dog_breed' and 'probability' from image_predictions_df dataframe
4. 'dog_breed' to lowercase for consistency in the image_predictions_df dataframe
5. columns we don't need, filled with 'Nan', 'in_reply_to_status_id', 'in_reply_to_user_id', from twitter_archive_df dataframe
6. we need only original tweets and no retweets; drop the columns 'retweeted_status_id','retweeted_status_user_id' and 'retweet_status_timestamp' from twitter_archive_df dataframe
7. datatypes inconsistency, for example : timestamp column datatype in twitter_archive_df dataframe is a string instead of datetime datatype
8. there are invalid dog names like : 'a', 'an' and 'the' in twitter_archive_df dataframe
9. rating needs to be calculated and the column needs to be created in twitter_archive_df dataframe
10. there are extremely high values for min and max statistical levels on twitter_archive_df and tweet_json_df dataframes
11. inconsistency on number of observations between the 3 datasets

Tidiness issues:

1.Dog stages: doggo, floofer, pupper, puppo, are spread through several columns, on twitter_archive_df.A dog_stage column can be created and deleted unnecessary ones.

2.The 3 datasets provided needs to be combined.

3. Data Cleaning

For the beginning of data cleaning I made copies for each dataframe and named them: twitter_archive_clean, image_predictions_clean and tweet_json_clean.

Quality issues

1. changed 'id' column name to 'tweet_id', from the tweet_json_clean dataframe
2. dropped unnecessary columns with probability and algorithms used, containing non-descriptive columns : 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog', 'img_num', from image_predictions_clean dataframe
3. .renamed 'p1', 'p1_conf' to 'dog_breed' and 'probability' from image_predictions_clean dataframe
4. changed 'dog_breed' to lowercase for consistency in the image_predictions_clean dataframe
5. dropped the columns we don't need, filled with 'Nan', 'in_reply_to_status_id', 'in_reply_to_user_id', from twitter_archive_clean dataframe
6. dropped the columns 'retweeted_status_id','retweeted_status_user_id' and 'retweet_status_timestamp' from twitter_archive_clean dataframe

7. changed timestamp column datatype from a string to datetime datatype in twitter_archive_clean dataframe
8. replaced invalid dog names like : 'a', 'an' and 'the' with ‘None’ in twitter_archive_clean dataframe

Tidiness issues:

1. created dog_stage column and deleted: doggo, floofer, pupper, puppo, on twitter_archive_clean.
2. merged the 3 clean datasets in one dataframe called df.